Claims

1. An apparatus for extracting information from a formatted document, comprising: an input unit (1) for inputting a formatted document; a unit (2) for analyzing the input formatted document and saving the particular typographic information; a unit (3) for identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc., a unit (4) for extracting the identified special character strings; and an output unit (5) for outputting the extracted character strings.

2. The apparatus for extracting information from a formatted document according to claim 1, wherein said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the typographic information of said formatted document when the typographic information of said character string is determined as a special typographic information.

3. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the font size of said character string is determined to be the biggest one among the surrounding character strings.

4. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the color and the font of said character string is determined to be a special one among the surrounding character strings.

5. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein said formatted document is HTML document, and said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the font of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

6. The apparatus for extracting information from a formatted document according to claim 1 or 2, wherein

said formatted document is HTML document, and said unit (3) for identifying special character strings determines a certain character string as a special one on the basis of the analyzing results with respect to said HTML document when the color of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

7. A method for extracting information from a formatted document, comprising the following steps ; inputting a formatted document, analyzing the input formatted document and saving the particular typographic information; identifying special character strings on the basis of the analysis result by means of the typographic information such as font size, character font, color, etc.; extracting the identified special character strings; and outputting the extracted character strings.

8. The method according to claim 8, wherein in the step of identifying special character string, a celtain character string is determined as a special one on the basis of the typographic information of said formatted document when the typographic information of said character string is determined as a special typographic information.

9. The method according to claim 7 or 8, wherein said formatted document is HTML document, and in the step of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the font size of said character string is determined to be the biggest one among the surrounding character strings.

10. The method according to claim 7 or 8, wherein said formatted document is HTML document, and in the step of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the color and the font of said character string is determined to be a special one among the surrounding character strings.

11. The method according to claim 7 or 8, wherein said formatted document is HTML document, and in the step of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the font of said character string is determined to be different from the surrounding character strings and said character string to be boldface.

12. The method according to claim 7or 8, wherein said formatted document is HTML document, and in the step

of identifying special character string, a certain character string is determined as a special one on the basis of the analyzing results with respect to said HTML document when the color of said character string is determined to be different from the surrounding character strings and said character string to be boldface.